Shironaam: Bengali News Headline Generation using Auxiliary Information

Abu Ubaida Akash[*§] **Mir Tafseer Nayeem**[*◊] **Faisal Tareque Shohan**[§] **Tanvir Islam**[‡]
[§]Ahsanullah University of Science and Technology [◊]University of Alberta
[‡]University of Hawaii at Manoa
akash.ubaida@gmail.com, mnayeem@ualberta.ca
faisaltareque@hotmail.com, tislam@hawaii.edu

Abstract

Automatic headline generation systems have the potential to assist editors in finding interesting headlines to attract visitors or readers. However, the performance of headline generation systems remains challenging due to the unavailability of sufficient parallel data for low-resource languages like Bengali and the lack of ideal approaches to develop a system for headline generation using pre-trained language models, especially for long news articles. To address these challenges, we present Shironaam, a large-scale dataset in Bengali containing over 240K news article-headline pairings with auxiliary data such as image captions, topic words, and category information. Unlike other headline generation models, this paper uses this auxiliary information to better model this task. Furthermore, we utilize the contextualized language models to design encoder-decoder model for Bengali news headline generation and follow a simple yet cost-effective coarse-to-fine approach using topic-words to retrieve important sentences considering the fixed length requirement of the pre-trained language models. Finally, we conduct extensive experiments on our dataset containing news articles of 13 different categories to demonstrate the effectiveness of incorporating auxiliary information and evaluate our system on a wide range of metrics. The experimental results demonstrate that our methods bring significant improvements (*i.e.,* 3 to 10 percentage points across all evaluation metrics) over the baselines[1]. Also to illustrate the utility and robustness, we report experimental results in few-shot and non-few-shot settings.

1 Introduction

News headlines can significantly affect the number of visitors and play a crucial part in the life-cycle of a news article (Murao et al., 2019). Therefore, representative and interesting headlines are arguably essential to any news document to grab the attention of potential readers (Mishra et al., 2021; Ao et al., 2021). Nowadays, online and printed news releases significantly increase the article's visibility, support, and context by using multimedia content. As a picture is worth a thousand words, digital assets such as images and videos are the go-to candidates for the thumbnails used in different social media, blogs, and many other platforms. The captions that go with the images or videos are equally significant as the actual content. Captions describing the images can clarify and enhance the image, optimize news articles for search engines, and improve the accessibility of the news for people with vision impairments[2].

Headline generation, given a news article, is a special case of abstractive summarization (Yamada et al., 2021), which involves sentence compression, syntactic reorganization, sentence fusion, and lexical paraphrasing (See et al., 2017; Gehrmann et al., 2018; Zhong et al., 2019; Nayeem et al., 2019; Nayeem and Chali, 2017b). Unlike text summaries, which often feature many or single long sentences to summarize a document's important concepts (Nayeem and Chali, 2017a), news headlines frequently have a single short catchy statement to grab the readers attention and entice them to read the story. Even though Bengali is the seventh most spoken language with approximately 337 million speakers worldwide[3] (Chakraborty et al., 2021; Chowdhury et al., 2021), generating quality headlines for a low-resource language such as Bengali is more challenging due to the unavailability of large-scale human-annotated dataset (Haque et al., 2016; Nayeem et al., 2018; Joshi et al., 2019).

[*] Equal contribution.

[1] Code, dataset, and model checkpoints: https://github.com/dialect-ai/BenHeadGen

[2] In this paper, we limit our focus to only captions to improve the news headlines. Using multimodal information for this task is left as possible future work.

[3] https://w.wiki/57

Contextualized language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 1907), T5 (Raffel et al., 2020) help improving several downstream tasks in NLP, such as summarization, question answering, and text classification. Unfortunately, these models suffer from a limitation as they can handle input sequences up to a certain limit (Sun et al., 2019). As a result, this limitation burdens some NLP tasks, especially where the input is necessarily long (Kitaev et al., 2020), such as transcript analysis of the phone calls, document topic prediction, news headline generation, etc. The most natural way to address this problem is to trim the input sequences to a maximum length. However, trimming the long input document is tricky, especially for headline generation. The news articles usually maintain coherence and relevant parts may be located at the bottom of the document, which may prevent models from generalizing well to positions beyond the cutoff point[4]. In this paper, we utilize topic words to retrieve important sentences as a context for the BERT model by following a simple yet cost-effective coarse-to-fine approach.

We present **Shironaam**, a large-scale abstractive Bengali news article dataset that includes over 240K professionally annotated headline-article pairings as well as auxiliary information such as image captions, topic words, and category information. Each sample can be represented as a tuple of (article, image caption, topic-words, category, headline). To the best of our knowledge, Shironaam is the first Bengali news article dataset incorporating auxiliary information and a benchmark for the news headline generation task. This corpus has the potential to authorize and encourage research on such a low-resource language, bringing technological advancements to a previously underserved community. Rather than the one-to-one mapping (*i.e.,* input is an article, and output is a headline) used in the earlier works (Takase et al., 2016; Zhang et al., 2018; Murao et al., 2019; Colmenares et al., 2019; Song et al., 2020; Li et al., 2021), we treat the headline generation task as a **three-to-one** mapping with the inputs being an image caption, a list of topic words, and an article where the output is a headline. Based on the transformer architecture, we utilize pre-trained language models for generating headlines and present a new concept of fusing image caption parallelly (Liu et al., 2020a) with the input article to support the three-to-one mapping and to encode long documents. We design and compare numerous input mechanism alternatives as part of the suggested strategy. Extensive experiments on our proposed dataset reveal that the suggested method is capable of generating high-quality news headlines (see Section F in the Appendix) and brings significant improvements over the state-of-the-art baselines across all evaluation metrics (Table 5).

Our main contributions can be summarized as follows:

- We provide **Shironaam**, a large-scale news headline generation dataset of a low-resource language *i.e.,* Bengali containing over 240K news headline-article pairings with auxiliary information such as image captions, topic words, and category information (Table 2). Also, this dataset can potentially be used for other tasks such as document categorization, news clustering, keyword identification, etc.
- We present a new concept of incorporating auxiliary information to model input with articles to improve the quality of headlines. We train an encoder-decoder model for this task almost from scratch, which utilizes pre-trained language model (Figure 1).
- We develop **BenSim**, an independent module for measuring the semantic similarity among Bengali sentences. We make use of the BenSim module and utilize topic words to encode long articles by following a simple yet effective approach (Figure 1(c)).
- To illustrate the utility and robustness, we also evaluate the performance with few-shot settings where the domains don't have enough training samples (Table 6).

2 The Shironaam Corpus

In this section, we present the first-of-its-kind corpus (we name it Shironaam) for news headline generation in Bengali like low-resource language. This includes auxiliary information in addition to the usual headline-article pairs. We explain the curation process involving raw data crawling, preprocessing, and cleaning.

[4] While Longformer (Beltagy et al., 2020) is a viable solution for this problem, it comes up with a high computational cost, and pre-trained models aren't available.

[5] https://en.wikipedia.org/wiki/Jaccard_index

Category	**Train**	**Valid**	**Test**	**Total**	**Jaccard (%)**
Entertainment	16,104	365	1095	17,565	13.56
National	117,566	2,664	7,994	128,226	24.60
Nature	467	10	31	510	23.66
International	30,558	692	2,078	33,329	18.09
Sports	17,635	399	1,199	19,235	17.82
Economy	6,447	146	438	7,032	39.37
Life-Health	6,356	144	432	6,933	17.83
Miscellaneous	1,599	36	108	1,744	11.71
Opinion	3,501	79	238	3,819	38.41
Politics	15,018	340	1,021	16,380	23.02
Edu-Career	4,008	90	272	4,372	53.58
Science-Tech	1,046	23	71	1,141	22.95
Religion	269	6	18	294	71.59
Total/Avg.	**220,574**	**4,994**	**15,012**	**240,580**	**28.94**

Table 1: Our headline generation dataset (**Shironaam**) distribution over **13** different domains. Jaccard scores[5] represent the similarities of each domain in between the image captions and headlines.

Features	**IndicNLG-BN**	**Shironaam (*ours*)**
Article	✓	✓
Headline	✓	✓
Category	✗	✓
Topic words	✗	✓
Image Caption	✗	✓
#Examples	142,731	240,580

Table 2: Feature-level comparison between IndicNLG-BN (2022) and **Shironaam** dataset (*ours*).

Dataset	**% of novel n-gram**			
	unigram	bigram	trigram	4-gram
Shironaam	26.59	66.12	82.71	86.49
IndicNLG BN	46.38	78.92	90.39	94.77

Table 3: Percentage (%) of novel n-grams between IndicNLG-BN (2022) and **Shironaam** dataset (*ours*).

2.1 Raw Data Crawling

We crawl around 900,000 raw data samples from seven famous Bengali newspapers (names in Section C in the Appendix) concentrating on certain criteria, such as headline, article, image caption, category, and topic words. Since each of the newspapers mentioned above has it's own professional authors and distinct writing style, we consider multiple sources to prevent the bias of a particular annotation style. To ensure content diversity, we also cover various domains from all the news dailies. The majority of the news samples are extracted from HTML bodies of the corresponding publications, while some are rendered using JavaScript. However, two of them (see in Appendix Section C) do not provide the archives on their websites; therefore, we collect the samples through their APIs.

2.2 Dataset Preprocessing

The overall crawled corpus contains a lot of noise, such as irrelevant details about the publisher and the date/time of the news in multiple formats, embedded advertisements, phrases from different languages (especially English), reference URLs, inconsistent bold sections, emoticons, extrinsic symbols, and various Unicode representations. Thus, we remove the date/time and the embedded items using regular expressions. To preserve only the Bengali texts, we construct a vocabulary of Bengali unit characters and perform character level matching in the article bodies and headlines. The English numbers, however, are retained since they are used frequently in regular Bengali texts.

The image captions sometimes include extra/irrelevant information (*e.g.,* ছবি [Picture], সংগৃহীত [Collected], ফাইল ছবি [File Image], রয়টার্স [Reuters], ইন্টারনেট [Internet], প্রতীকী ছবি [Symbolic Image], *etc.*)[6], which are common in any news article. Thus, we identify these repetitive words using a simple frequency-based approach over all the samples and remove them from the image captions. Furthermore, we discard the samples whose captions are smaller than four words in length; from our manual inspection, we observed that these words often describe the named entities present in the image, such as name, location, date/time, etc.

Different newspapers use different names to categorize their contents. Consequently, each domain is represented with different category names in all the news dailies. For extracting the categories, we map them with their corresponding representative domains and label each domain with it's relevant names. For instance, national, whole-country, city-news, country, capital, city-roundup, south-city, *etc.* are distinct categorical terms, but they can be grouped easily under the *national* domain. Table 1 shows the distribution of the final domains in the Shironaam corpus. We use `sbnltk`[7] for tokenizing the documents into sentences. Finally, we discard the samples where any of the information (*i.e.*, headline, article, or image caption) is missing.

2.3 Dataset Statistics

After preprocessing the raw corpus, we have 240,580 news samples as a tuple of (headline, article, image caption, topic words, category). To en-

[6] The square brackets contain the English translations.
[7] `https://pypi.org/project/sbnltk`

Dataset	Article	Headline	Image Caption	Topic Words
Average number of words				
Shironaam	252.01	6.53	6.80	3.21
IndicNLG BN	199.83	10.03	-	-
Average number of sentences				
Shironaam	20.05	1.00	1.04	-
IndicNLG BN	15.19	1.19	-	-
Vocabulary size				
Shironaam	605,750	76,732	87,644	-
IndicNLG BN	614,374	65,553	-	-

Table 4: Quantitative statistics compared to IndicNLG-BN (2022) and our proposed dataset **Shironaam**.

sure a balanced distribution, we maintain the ratio of (92% - 220,574), (2% - 4994), and (6% - 15,012) samples from all the categories to construct the train, validation, and test set, respectively (see Table 1). We compare our corpus with the only available benchmark, IndicNLG (Kumar et al., 2022), for the news headline generation task in Bengali. Since IndicNLG covers multiple languages, we just keep the Bengali (BN) language portion for comparison. Table 2 provides a high-level summary of both datasets.

Our Shironaam corpus establishes a new benchmark in terms of the corpus size compared to IndicNLG (Kumar et al., 2022). It is important to note that our corpus also contains auxiliary information such as image captions, topic words, and article categories. Moreover, this can be used not only in headline generation tasks but also in some other tasks such as document categorization, news clustering, keyword identification, etc. To measure the abstractiveness, in Table 3, we calculate the percentage of novel n-grams in reference headlines that are not present in the article. Table 3 shows that the novelty level increases with the number of grams, and the average scores are comparable to the IndicNLG (Kumar et al., 2022). A quantitative statistics presented in Table 4 demonstrates that our Shironaam corpus contains more compressed headlines against lengthier articles compared to the IndicNLG, both in terms of words and sentences. This highly compressed nature of the headlines makes the task of headline generation in low-resource language more challenging. In addition, the vocabulary size of our articles is comparable with IndicNLG, whereas we get a larger number of vocabularies in our headlines (see Table 4).

Therefore, the Shironaam corpus comprises a diverse range of headline styles and provides the largest collection of Bengali news articles. Moreover, it is the first benchmarking dataset in such a low-resource language that includes auxiliary information in addition to the headline-article pairs. We hope it will motivate further study and serve as a baseline for future works on this task for this low-resource language.

3 News Headline Generation

3.1 Task

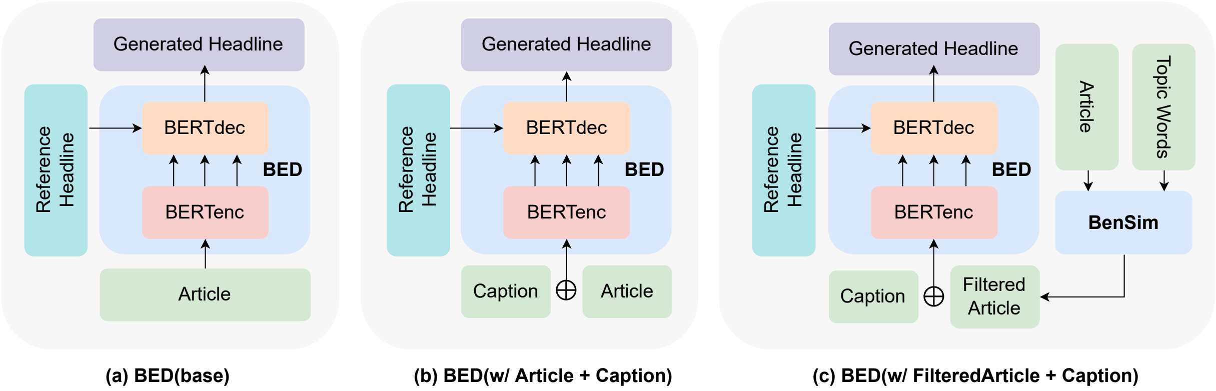
We establish a new concept of incorporating auxiliary information in order to generate high-quality headlines in Bengali, a low-resource language. In the context of this generation task, we assume that a) we have enough data with auxiliary information to train a headline generation model in Bengali language (can be referred to Shironaam corpus); b) the auxiliary information refers to the image captions and topic words used in tagging documents; c) we have access to a module that filters a document based on the contextual similarity with a list of topic words (we refer BenSim in Section 3.3). The task can be formalized as follows. Given article $\mathcal{A}$, image caption $\mathcal{C}$, and a set of topic words $\mathcal{T}$ as input, our goal is to generate high-quality headline $\mathcal{H}$ for the corresponding news article.

3.2 Approach

To carry out the idea, we need several benchmarks to compare with and evaluate our proposed hypothesis. But, no SOTA benchmark is available for this task in Bengali language[8], except the IndicBART (Dabre et al., 2022). So, we set multiple baselines (Section 4.2), both of extractive and abstractive types, that take article $\mathcal{A}$ as input and generate corresponding headline $\mathcal{H}$ as output. We follow LEAD-1 and EXT-ORACLE approaches among the extractive types, whereas from the abstractive types, we initialize an encoder-decoder model for Bengali language with a pre-trained encoder-only checkpoint to skip the costly pre-training (Rothe et al., 2020) [9]. To train the encoder-decoder model (BED), we use BanglaBERT (Bhattacharjee et al., 2022a) as the encoder checkpoint. Additionally, we utilize other pre-trained models (*i.e.,* BanglaT5

[8] We did not consider the extreme summarization models since the style of a headline and a single-line summary is completely different.

[9] We refer the interested readers to Appendix (Section A) for necessary background.



(a) BED(base)

(b) BED(w/ Article + Caption)

(c) BED(w/ FilteredArticle + Caption)

Figure 1: Graphical illustration of our proposed headline generation models **(a)** BERT-based encoder-decoder (*baseline*) **(b)** Our model incorporating image caption with the article **(c)** Our model uses **BenSim** to extract important sentences (*a.k.a.,* filtered article) based on topic words and incorporates filtered article with the image caption.

(Bhattacharjee et al., 2022b), IndicBART (Dabre et al., 2022)) based on transformer architecture (Vaswani et al., 2017). After comparing all the baselines (see Section 4.2), we select the best performing one for further ablations. Experimental results (in Table 5) reveal that BED model outperforms other baselines, even though the fine-tuned BanglaT5 (Bhattacharjee et al., 2022b) scores competitively.

3.3 BERT-based Encoder-Decoder (BED)

A BED model consists of an encoder that has been initialized with BERT, termed as BERTenc, coupled with a decoder that has also been initialized with BERT, which we call BERTdec. The initializing point for each weight's calculation is a public BERT checkpoint. The only variable initialized at random is the encoder-decoder attention (Rothe et al., 2020).

Article Only In Figure 1(a), we implement the basic version *i.e.,* BED (base) model, which takes word tokens of an article as a sequence of inputs $\mathcal{A}_{1:n}$ and describes a conditional distribution of target vectors $\mathcal{H}_{1:l}$ of variable length **l**, in our case, generated words for headline:

$$p_{\theta_{BERTenc},\theta_{BERTdec}}(\mathcal{H}_{1:l}|\mathcal{A}_{1:n}). \quad (1)$$

The input sequence $\mathcal{A}_{1:n}$ is sent to the BERTenc component, which then converts it into a sequence of hidden states, $\overline{\mathcal{A}}_{1:n}$. The mapping can be defined as:

$$f_{\theta_{BERTenc}} : \mathcal{A}_{1:n} \to \overline{\mathcal{A}}_{1:n}. \quad (2)$$

The BERTdec component will simulates the conditional probability distribution of the target vector sequence $\mathcal{H}_{1:l}$, assuming that the sequence of encoded hidden states $\overline{\mathcal{A}}_{1:n}$ has been provided:

$$p_{\theta_{BERTdec}}(\mathcal{H}_{1:l}|\overline{\mathcal{A}}_{1:n}). \quad (3)$$

Bayes' rule lets us turn this distribution into a product of the conditional probability distribution of the target vector $\mathbf{h}_i$, given the encoded hidden states $\overline{\mathcal{A}}_{1:n}$ and all the previous target vectors $\mathcal{H}_{0:i-1}$:

$$\begin{aligned} &p_{\theta_{BERTdec}}(\mathcal{H}_{1:l}|\overline{\mathcal{A}}_{1:n}) \\ &= \prod_{i=1}^{n} p_{\theta_{BERTdec}}(\mathbf{h}_i|\mathcal{H}_{0:i-1}, \overline{\mathcal{A}}_{1:n}). \end{aligned} \quad (4)$$

All preceding target vectors $\mathcal{H}_{0:i-1}$ and the encoded hidden state sequence $\overline{\mathcal{A}}_{1:n}$ are mapped to the logit vector $\mathcal{V}_i$ by the BERTdec. The next step is to run the softmax operation on the logit vector $\mathcal{V}_i$. This helps to define the conditional distribution $p_{\theta_{BERTdec}}(\mathbf{h}_i|\mathcal{H}_{0:i-1}, \overline{\mathcal{A}}_{1:n})$ by making sure that the distribution of the target vector $\mathbf{h}_i$ depends on the distributions of all previous target vectors $\mathbf{h}_0, \dots, \mathbf{h}_{i-1}$:

$$p_{\theta_{BERTdec}}(\mathbf{h}_i|\mathcal{H}_{0:i-1}, \overline{\mathcal{A}}_{1:n}) = \mathbf{Softmax}(\mathcal{V}_i). \quad (5)$$

The first target vector $\mathbf{h}_0$ is going to be represented by a unique BOS vector that is referred to as the "beginning-of-sentence". After the conditional distribution $p_{\theta_{BERTdec}}(\mathbf{h}_i|\mathcal{H}_{0:i-1}, \overline{\mathcal{A}}_{1:n})$ has been set, the output can be made in an auto-regressive way. This makes it possible to define a mapping between an input sequence $\mathcal{A}_{1:n}$ and an output sequence $\mathcal{H}_{1:l}$ at the time of inference.

Fusing Article and Image Caption In order to explore more ways to improve the quality of the generated headlines, we employ BED (w/ Article + Caption) model, which incorporates image caption $\mathcal{C}_{1:m}$ with the corresponding article $\mathcal{A}_{1:n}$ as in Figure 1(b), where **m** << **n**, and passes them through the BERTenc using parallel-fusion (Liu et al., 2020a) mechanism:

$$\mathcal{K}_{1:r} = \mathcal{C}_{1:m} \oplus \mathcal{A}_{1:n}, \tag{6}$$

$$f_{\theta_{BERTenc}} : \mathcal{C}_{1:m}, \mathcal{A}_{1:n} \rightarrow \overline{\mathcal{K}}_{1:r}. \tag{7}$$

Here, $\mathcal{K}_{1:r}$ denotes the model input sequence, where **r** represents the new input sequence length, and $\oplus$ is concatenation operator separated by a special token. The sequence of hidden states $\overline{\mathcal{K}}_{1:r}$ are then processed through the BERTdec likewise the Shironaam(base) model and the headline is generated as output:

$$p_{\theta_{BERTenc}, \, \theta_{BERTdec}}(\mathcal{H}_{1:l} | \mathcal{C}_{1:m}, \mathcal{A}_{1:n}). \tag{8}$$

However, the image caption may not always serve the full context if the news article becomes too long for BERTenc. Moreover, the image caption length is generally much smaller than the news article length. Thus, the impact of using image caption as a context is less sensitive for lengthier articles.

Bengali Sentence Similarity (BenSim) Since many of the news articles' lengths exceed the input sequence limit that BERTenc can process, we therefore, utilize the sequence length by ensuring all the relevant sentences are present in the limited input sequences. To ensure the extraction of relevant sentences, we develop BenSim module[10], a tool for measuring semantic similarity between Bengali sentences utilizing BERT embeddings. It takes news article $\mathcal{A}_{1:n}$ and corresponding topic words $\mathcal{T}_{1:k}$ as input for getting most of the contextual sentences and employs pre-trained bangla-bert-base (Sarker, 2020) model on both of the input sequences to generate the contextualized encoded representations. After performing mean pooling operation, cosine similarity (Singhal, 2001) is then applied to the encoded sequences to get the similarity score. After measuring the similarities between the topic words and input sentences, a filtered article $\mathcal{A'}_{1:r}$ is returned as output, which is then fused parallelly with the image caption $\mathcal{C}_{1:m}$ and sent to the model input. Finally, the BED (w/ FilteredArticle + Caption) model produces a headline after processing the fused input.

4 Experiments and Benchmarks

In this section, we set a new benchmark for Bengali news headline generation using Shironaam corpus and compare it with the other state-of-the-art baselines. After a clean comparison, we perform two ablation experiments on the superior base model. Then, we analyze the performance gap between the baselines and the ablations and further evaluate the best model on news domains with a few samples (few-shot). Finally, after proper analysis, we seek to find out the answers to the following research questions:

- **RQ#1:** Can we use auxiliary information (*e.g.,* image caption and topic words) to improve the performance of the headline generation?
- **RQ#2:** Which domain(s) benefit from the auxiliary information in few-shot and non-few-shot settings?

4.1 Implementation Details

We utilize the encoder-decoder paradigm[11] of HuggingFace, where pre-trained BanglaBERT (Bhattacharjee et al., 2022a)[12] is used to initialize both of the weights of encoder and decoder. Before proceeding to tokenization, we perform sentence normalization, introduced in Hasan et al. (2020). For tokenization, we use the pre-trained tokenizer[12] that comes with the model. All the hyper-parameters used for training and decoding are presented in Section B in the Appendix.

Evaluation Metrics We compare the performance with the following baselines across several evaluation metrics presented in Section D in the Appendix.

4.2 Baselines

LEAD-1 LEAD-1 is a commonly used baseline for setting the lower bound of news headline generation task (Kumar et al., 2022; Narayan et al., 2018). It also indicates the degree of positional biasness of article body sentences in generating headlines. We pick the article's first sentence as the system headline and compare it with the original headline to generate the LEAD-1 scores.

[10] https://github.com/dialect-ai/BenSim

[11] Encoder-Decoder models documentation

[12] BanglaBERT usage (HuggingFace)

Models	ROUGE			BLEU			BERT	METEOR
	R-1	R-2	R-L	BLEU Score	Brevity Penalty	Length Ratio	Score	Score
Baselines								
LEAD-1 (Extractive)	30.50	13.86	28.00	5.65	97.71	2.48	74.63	29.90
EXT-ORACLE (Extractive)	39.92	22.89	37.28	9.17	97.16	2.30	77.16	39.65
IndicBART (mBART)	28.76	12.65	27.11	15.03	99.91	1.14	74.95	20.39
BanglaT5 (mT5)	44.13	23.03	42.12	13.05	91.33	1.15	80.13	34.65
Our Ablations								
BED Base (BERT2BERT)	44.22	24.18	42.28	22.06	94.47	0.94	80.53	34.16
-w/ Article + Caption	51.62	33.62	49.94	31.39	96.02	0.96	82.93	42.57
-w/ FilteredArticle + Caption	**52.19**	**34.27**	**50.31**	**31.80**	**98.57**	**0.99**	**83.10**	**43.52**

Table 5: Performance on `Shironaam (test)` corpus compared to the baselines (Section 4.2) and the results of our ablation study (see Appendix Section E for validation scores) across various automatic evaluation metrics, where **bold-faced** scores indicate superior performance.

EXT-ORACLE On the other hand, `EXT-ORACLE` can be considered as the upper bound of generating headlines by an extractive approach (Kumar et al., 2022; Narayan et al., 2018). We implement this baseline on the `Shironaam (test)` corpus by aligning a sentence from the input article with the reference headline based on the ROUGE-2 metric.

IndicBART Kumar et al. (2022) releases a multilingual model, which is fine-tuned on `IndicBART` (Dabre et al., 2022) checkpoint for the headline generation task focusing on Indic languages including Bengali. `IndicBART` is a sequence-to-sequence multilingual pre-trained model (Dabre et al., 2022) based on the mBART (Liu et al., 2020b) architecture.

BanglaT5 We fine-tune `BanglaT5` (Bhattacharjee et al., 2022b), a sequence-to-sequence transformer model based on mT5 (Xue et al., 2021) architecture for Bengali language, on the `Shironaam (train)` corpus for the headline generation task. For a fair comparison, we maintain the same hyper-parameters.

BED (base) Model We implement the model (article only), illustrated in Figure 1(a) on the `Shironaam (train)` corpus to make the baseline. We utilize 220,500 news samples from the train set to train the `BED` model, which takes the article only as input and generates a headline as output. The evaluation result on the `Shironaam (test)` set is shown in Table 5, which is a new benchmark for the Bengali headline generation task. In the following experiments, we utilize the auxiliary information with the same hyper-parameter settings to generate better-quality headlines.

4.3 Ablation Experiments

BED (w/ Article + Caption) Model As per the demonstration in Figure 1(b), the image caption is incorporated with the input article. This leads to a much improved result across all evaluation metrics compared to article only model (a.k.a., `BED (base)`) as shown in Table 5.

BED (w/ FilteredArticle + Caption) Model Since the utilization of image caption in model input gives better results, therefore we further enrich the inputs by incorporating topic words. We use topic words in filtering the longer articles through `BenSim` rather than using them directly to the input, as shown in Figure 1(c). First, we set a threshold value (40 in our case) for `BenSim` to extract the number of top semantically similar sentences. `BenSim` maintains the relative appearance order of the sentences in the original article to construct the corresponding filtered article. To fix the number of sentences in a filtered article, we consider the maximum use of the number of tokens `BED` model can afford *i.e.* 512. Fusing filtered article with image caption achieves the best results across several evaluation metrics as shown in Table 5 (also see Appendix Section F for the generation quality).

4.4 Discussions

Result Analysis Table 5 shows that the `LEAD-1` baseline performs inadequately on the `Shironaam (test)` corpus. More specifically, the ROUGE-2 and BLEU scores and the length ratio indicate that the original headlines are more abstractive in nature, and the first sentence of an article does not contain sufficient information for generating a headline. Unlike `LEAD-1`, comparatively higher ROUGE scores are obtained by using

EXT-ORACLE, but at the same time, BLEU score gain is lower. This trade-off indicates that the reference headlines consist of the subset of words present in the sentences selected by EXT-ORACLE. However, because of the concise nature of news headlines, this approach does not fit well but can be considered a strong baseline for other models. Among the abstractive types, IndicBART performs poorly on the Shironaam (test) corpus and even is unable to beat the weak LEAD-1 baseline, let alone EXT-ORACLE. On the other hand, the fine-tuned BanglaT5 yields a good score for this task. Although the generated results are slightly lengthier than the reference ones, they can be considered a strong baseline. The BED (base) model provides the best performance in terms of ROUGE, BLEU, and BERT scores. So, we consider it the strongest baseline and look for further ablations.

To this end, Table 5 shows that the best baseline is outperformed by our proposed technique of modeling input using auxiliary data. We want to emphasize that we use image caption and topic words purely as auxiliary data. While collecting the data from various news portals, we observe that it is very common to include images to help support and communicate the story and image captions are a crucial part of it that only describe the referred image. Although image captions are mostly correlated with the corresponding article in terms of context, we argue that they are not headlines. First, there is not much overlap in terms of Jaccard similarity measured between image captions and headlines (as from Table 1 we have approximately 29% overlap across different categories). Second, headlines differ from image captions in terms of styles and content.

Image captions usually give the model some signal on which parts of the document model need to attend more. Hence, as a result of combining image caption with article, BED (w/ Article + Caption) model improves the performance by about 3 (BERT score) to 10 (ROUGE-2 score) percentage points and produces more human-like headlines. Moreover, it often begins generating sentient headlines that are more abstract and profound than the reference ones. The BED (w/ FilteredArticle + Caption) model performs slightly better than the previous ablation. Since, there are fewer lengthier articles in the Shironaam corpus, the variations in the scores of the two ablation models are rather small. We observe that when we include filtered articles led by relevant topic words in the model's input, it begins to learn faster than the model without topic words (demonstrated in Figure 2). Therefore, the differences between the scores of two ablation models will increase with the number of lengthier articles.

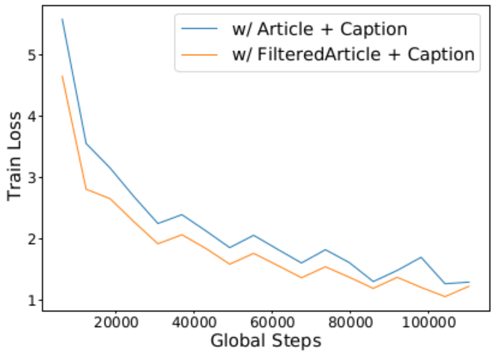


Figure 2: Train Loss vs. Global Steps for our ablations.

So, following the question **RQ#1**, we may conclude from the preceding discussion that auxiliary information definitely aids in creating better headlines. Although we achieve superior performance compared to the state-of-the-art baselines across several evaluation metrics, these quantitative measures can not determine the generation quality. Therefore, we present generated samples from our model categorized into several abstractive types (see Section F in Appendix). We leave the human evaluation of our generated samples as one of the future works.

Domain Specific Analysis We evaluate our proposed BED (w/ FilteredArticle + Caption) [denoted as **BED (FA+C)**] model on individual domains by comparing with a base model and to answer **RQ#2**. We also observe the performance of the presented model on the domains with fewer samples (few-shot). We employ two baselines here: BED (base) and BanglaT5 (Bhattacharjee et al., 2022b) [denoted as **BNT5**]. Although, the BNT5 has demonstrated competitive performance, Table 6 shows that BED (base) model performs better on the maximum number of domains. To calculate the exact performance gap, we maintain a uniform baseline *i.e.,* BED (base) to compare with the proposed model. For the few-shot observation, based on the number of training samples, we split the domains into two folds (see Table 6). The *Few-Shot* domains contain less than 6500 train samples, whereas rest of the domains are considered as *Non-Few-Shot*.

Category	R-1			R-2			R-L		
	BED (base)	BNT5	BED (FA+C)	BED (base)	BNT5	BED (FA+C)	BED (base)	BNT5	BED (FA+C)
Non-Few-Shot Domains									
National	48.03	47.33	55.84	27.29	25.83	37.88	46.06	45.37	53.95
International	44.44	46.04	50.47	22.92	23.08	29.96	42.02	43.49	48.13
Sports	30.14	33.46	39.20	11.57	13.43	20.40	28.75	31.59	37.33
Entertainment	33.05	32.99	35.14	15.07	14.32	16.64	31.26	31.33	33.44
Politics	49.28	49.66	**57.16**	28.80	27.32	**39.73**	47.53	47.68	**55.73**
Few-Shot Domains									
Economy	38.95	40.03	60.32	18.81	19.74	45.85	36.44	37.62	58.53
Life-Health	35.87	39.20	44.97	17.61	19.78	27.21	33.90	37.38	43.08
Edu-Career	50.57	51.12	71.55	31.92	30.82	59.54	48.05	48.82	70.48
Opinion	16.11	15.82	44.53	4.69	5.24	36.63	15.82	15.44	44.25
Miscellaneous	33.64	34.92	35.29	16.16	17.98	17.41	30.48	32.82	31.87
Science-Tech	41.82	44.14	51.03	19.54	22.61	31.20	39.30	41.82	48.49
Nature	36.07	37.89	46.54	15.78	16.65	30.07	34.84	35.79	45.53
Religion	27.29	35.48	**72.10**	12.28	19.63	**62.05**	26.96	34.42	**72.14**

Table 6: Performance of our proposed model BED(FA+C) compared to baseline BED(base) and BNT5 (2022b) across different domains. Shaded grey region indicates superior performance compared to baselines and **bold-faced** and underlined scores indicate comparably best and worst domains, respectively.

Table 6 demonstrates that our proposed model improves the scores by a satisfactory margin of almost all the domains except *Entertainment* and *Miscellaneous*. These two categories get comparatively lower scores. The majority of headlines in the *Entertainment* domain are casual and clickbait-style and do not maintain the identical nature of a particular domain. We argue that the discrepancy, in this case, decreases the scores. The *Miscellaneous* domain is comprised of different sorts of randomness containing articles of various domains. Therefore, it is anticipated that this genre will get a lower score. Table 6 shows that our proposed model maintains consistent performance when there are few samples to train.

5 Related Works

Headline generation is an under-explored subtask of abstractive summarization, particularly in languages with limited resources. For the English language, an attention-based neural network has been proposed by Rush et al. (2015) for abstractive sentence summarization. The authors propose a model that utilizes a recurrent neural network (RNN) and an attention mechanism to summarize input sentences into a compact summary. Takase et al. (2016) build an AMR encoder for headline creation based on an encoder-decoder architecture. Using a dual-attention seq2seq model, Zhang et al. (2018) proposes a way for question headline development. In limited resource settings, Tilk and Alumäe (2017) pretrain a neural encoder and decoder model to enhance headline generation outputs. A sentence encoder, a gate network for sentence selection, and a headline decoder are the three stages of Zhou et al. (2017)'s headline generation approach. Tan et al. (2017) proposes a coarse-to-fine strategy that extracts the most important sentences before generating the headlines based on the context. For headline generation, Kumar et al. (2022) have released the IndicNLG, a collection of multilingual datasets. However, they do not provide any additional attributes besides the headline-article pairs. In summary, the majority of the past works for generating headlines primarily used the article content to generate headlines.

6 Conclusion and Future Work

In this paper, we contributed a large-scale dataset (*a.k.a.,* **Shironaam**) with auxiliary information such as image captions, topic words, and category for Bengali news headline generation. We employ contextualized language models to incorporate such auxiliary information and proposed a simple yet effective solution to encode long articles using topic words. Experimental results demonstrate the superiority of our approach across different domains and settings. We anticipate that our efforts will motivate the community to expand the scope of headline generation tasks beyond English, particularly for a low-resource language like Bengali. Our future work will look into incorporating auxiliary information to support more languages and build a language-agnostic model.

Limitations

Our model relies on auxiliary information such as image captions and topic words to achieve superior performance. However, it is quite common to include images and extra information (*e.g.*, topic words) to increase the article's visibility, support, and context. Also, our base model without auxiliary information demonstrates improved performance compared to the well-established and state-of-the-art baselines. Another limitation we observed that our model did not perform as well as for the Miscellaneous and Entertainment categories compared to the other 11 different categories because of the clickbaity nature of these categories. Finally, our headline generation model works only for Bengali, a widely spoken but low-resource language. Still, this idea of using auxiliary information to improve headline generation performance can easily be extendable for many languages.

Ethics Statement

We considered some ethical aspects while scraping the data. We requested data at a reasonable rate without any intention of a DDoS attack. Moreover, for each website, we read the instructions listed in robots.txt to check whether we can crawl the intended content. We tried to minimize offensive texts in the data by explicitly crawling the sites where such contents are minimal. Further, we removed the Personal Identifying Information (PII) such as name, phone number, email address etc from the corpus.

Acknowledgements

We thank all the anonymous reviewers for their valuable feedback and constructive suggestions for improving this work. Mir Tafseer Nayeem is supported by a Huawei Doctoral Scholarship.

References

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022a. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2022b. Banglanlg: Benchmarks and resources for evaluating low-resource natural language generation in bangla. *CoRR*, abs/2205.11081.

Susmoy Chakraborty, Mir Tafseer Nayeem, and Wasi Uddin Ahmad. 2021. Simple or complex? learning to predict readability of bengali texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12621–12629.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Radia Rayan Chowdhury, Mir Tafseer Nayeem, Tahsin Tasnim Mim, Md. Saifur Rahman Chowdhury, and Taufiqul Jannat. 2021. Unsupervised abstractive summarization of Bengali text documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2612–2619, Online. Association for Computational Linguistics.

Carlos A Colmenares, Marina Litvak, Amin Mantrach, Fabrizio Silvestri, and Horacio Rodríguez. 2019. Headline generation as a sequence prediction with conditional random fields. In *Multilingual Text Analysis: Challenges, Models, And Approaches*, pages 201–243. World Scientific.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL*

2022, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Md. Majharul Haque, Suraiya Pervin, and Zerina Begum. 2016. Enhancement of keyphrase-based approach of automatic bangla text summarization. In *2016 IEEE Region 10 Conference (TENCON)*, pages 42–46.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 211–219, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M Khapra, and Pratyush Kumar. 2022. Indicnlg suite: Multilingual datasets for diverse nlg tasks in indic languages. *arXiv preprint arXiv:2203.05437*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ping Li, Jiong Yu, Jiaying Chen, and Binglei Guo. 2021. Hg-news: News headline generation based on a generative pre-training model. *IEEE Access*, 9:110039–110046.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Dayiheng Liu, Yeyun Gong, Yu Yan, Jie Fu, Bo Shao, Daxin Jiang, Jiancheng Lv, and Nan Duan. 2020a. Diverse, controllable, and keyphrase-aware: A corpus and method for news multi-headline generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6241–6250, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. 1907. Roberta: A robustly optimized bert pretraining approach. arxiv 2019. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Prakhar Mishra, Chaitali Diwan, Srinath Srinivasa, and G. Srinivasaraghavan. 2021. Automatic title generation for text with pre-trained transformer language model. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 17–24.

Kazuma Murao, Ken Kobayashi, Hayato Kobayashi, Taichi Yatsuka, Takeshi Masuyama, Tatsuru Higurashi, and Yoshimune Tabuchi. 2019. A case study on neural headline generation for editing support. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 73–82, Minneapolis, Minnesota. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Mir Tafseer Nayeem and Yllias Chali. 2017a. Extract with order for coherent multi-document summarization. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, pages 51–56, Vancouver, Canada. Association for Computational Linguistics.

Mir Tafseer Nayeem and Yllias Chali. 2017b. Paraphrastic fusion for abstractive multi-sentence compression generation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2223–2226, New York, NY, USA. Association for Computing Machinery.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2019. Neural diverse abstractive sentence compression generation. In *Advances in Information Retrieval*, pages 109–116, Cham. Springer International Publishing.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understading.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Amit Singhal. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24:35–43.

Yun-Zhu Song, Hong-Han Shuai, Sung-Lin Yeh, Yi-Lun Wu, Lun-Wei Ku, and Wen-Chih Peng. 2020. Attractive or faithful? popularity-reinforced learning for inspired headline generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8910–8917.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings*, page 194–206, Berlin, Heidelberg. Springer-Verlag.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on Abstract Meaning Representation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1054–1059, Austin, Texas. Association for Computational Linguistics.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4109–4115. AAAI Press.

Ottokar Tilk and Tanel Alumäe. 2017. Low-resource neural headline generation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 20–26, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Kosuke Yamada, Yuta Hitomi, Hideaki Tamori, Ryohei Sasano, Naoaki Okazaki, Kentaro Inui, and Koichi Takeda. 2021. Transformer-based lexically constrained headline generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4085–4090, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, Huanhuan Cao, and Xueqi Cheng. 2018. Question headline generation for news articles. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM ’18, page 617–626, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what’s next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058, Florence, Italy. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Vancouver, Canada. Association for Computational Linguistics.

A Preliminaries

Abstractive text summarization (Rush et al., 2015; See et al., 2017; Zhang et al., 2020) was considerably more challenging before the development of sequence-to-sequence (seq2seq) models (Cho et al., 2014; Sutskever et al., 2014) and recent advances in transformer-based models (Vaswani et al., 2017; Devlin et al., 2019) due to a lack of sufficient datasets. Many text-summarizing applications are still hindered by the lack of suitable datasets, particularly for low-resource languages (Joshi et al., 2019). After being presented in Vaswani et al. (2017), models based on transformer architectures have been proven to perform better on sequence-to-sequence tasks than decoder-only language models *e.g.* Raffel et al. (2020). In its most basic form, an encoder-decoder model comprises a stand-alone encoder, like BERT (Devlin et al., 2019), and a stand-alone decoder model, like GPT2 (Radford et al., 2019). It has been demonstrated that the huge pre-trained encoder-decoder models may considerably improve performance on a range of sequence-to-sequence tasks Lewis et al. (2020); Raffel et al. (2020). On the other hand, pre-training encoder-decoder models are very expensive to build since the models require a lot of computational resources.

Rothe et al. (2020) introduces the encoder-decoder model using pre-trained encoder and/or decoder-only checkpoints (such as BERT (Devlin et al., 2019) and GPT2 (Radford et al., 2019)) to avoid the time-consuming pre-training process. According to Rothe et al. (2020), these encoder-decoder models can do well as large pre-trained encoder-decoder models like T5 (Raffel et al., 2020) and Pegasus (Zhang et al., 2020) on different sequence-to-sequence tasks at a fraction of the training cost.

B Hyper-parameters, Training, and Decoding

All the BED models (Figure 1) are trained almost from scratch by maintaining uniform hyper-parameters and trained for 110,250 global steps with the learning rate 5e-5, and batch size 12. We save the best checkpoint by ensuring the lowest validation loss. We use AdamW (Loshchilov and Hutter, 2019) for optimizing the loss with default linear warmup. The maximum lengths of encoder and decoder are limited to 512 and 32 tokens, respectively. Each of the BED models is trained on a single NVIDIA Tesla P100 GPU and trained for approximately 33 hours, which takes almost 5 hours 30 minutes per epoch. The total number of trainable parameters is 249,044,480.

Decoding When validating and testing, we use beam search algorithm (Sutskever et al., 2014) with 4 beams to generate headlines. The maximum and minimum lengths used in generating the headlines are 16 and 4, respectively. We use ‘early stopping’ to stop the beam search when at least 4 sentences are finished per batch. The ‘no-repeat n-gram size’ is set to 2, where the exponential penalty to the length is 1.2. Regarding vocabulary size, we use the fixed 32,000 vocabularies from the encoder.

C Data Sources

Newspaper	URL
Prothom Alo	www.prothomalo.com
Naya Diganta	www.dailynayadiganta.com
Ajker Patrika	www.ajkerpatrika.com
Bangladesh Protidin	www.bd-pratidin.com
Samakal	www.samakal.com
Bhorer Kagoj	www.bhorerkagoj.com
Dhaka Tribune	www.dhakatribune.com

Table 7: List of Bengali newspapers to form the Shironaam corpus with their corresponding URLs. Samples from the *italic-faced* newspapers were crawled through their APIs.

D Evaluation Metrics

We evaluate the predicted headlines with some automatic metrics used for generation tasks. The generation quality is measured with the ROUGE (Lin, 2004) F1 score[13]. ROUGE-1 and ROUGE-2 measure informativeness, where fluency is measured by the longest common subsequence (ROUGE-L). We include BLEU (Papineni et al., 2002) score which indicates the similarity between reference and predicted sentences by comparing the overlap within tokens[14]. Brevity penalty and length ratio are shown to justify the BLEU score. The contextual similarity between the generated and reference headline is measured using F1 BERT score (Zhang* et al., 2020)[15], where the correlation between them is reported by METEOR score (Banerjee and Lavie, 2005)[16]. We use the available open-source implementations for the above metrics.

[13] ROUGE (multilingual)

[14] BLEU (HuggingFace)

[15] BERTScore (HuggingFace)

[16] METEOR (HuggingFace)

E Validation Results

	Train Loss	Valid Loss	ROUGE			BLEU		METEOR
			R-1	R-2	R-L	Bleu Score	Length Ratio	Score
a)	1.0892	2.4332	44.51	23.56	42.38	20.39	0.95	34.27
b)	1.5083	2.1227	49.59	30.53	47.79	27.63	0.98	40.59
c)	1.2199	2.0836	49.77	31.42	48.05	28.70	0.99	40.73

Table 8: All the scores are reported for BED model with ablations on Shironaam (valid) corpus. The labels indicate the ablations of BED model: **a)** Base, **b)** Article + Caption, **c)** FilteredArticle + Caption. Only the *Train Loss* is measured on the training set and kept for comparison with the *Valid Loss*.

F Generated Headlines

Generated headlines on **Shironaam** (*test*) corpus across all the categories are presented in Table 9.

Category	Headline		Type
Economy	GH	চিনি আমদানিতে শুল্ক কমালো সরকার	**Inserted**
	ET	The government reduced the duty on sugar import	
	RH	চিনি আমদানিতে শুল্ক কমলো	
	ET	Import duty on sugar reduced	
Edu-Career	GH	সহকারী জজ নিয়োগের লিখিত পরীক্ষার সূচি প্রকাশ	**Matched**
	ET	Release of written test schedule for the appointment of Assistant Judge	
	RH	সহকারী জজ নিয়োগের লিখিত পরীক্ষার সূচি প্রকাশ	
	ET	Release of written test schedule for the appointment of Assistant Judge	
Entertainment	GH	"আমি বিয়ে করব না, দেখি কে আমাকে বিয়ে করে"	**Swapped**
	ET	“I will not marry, let’s see who marries me”	
	RH	"আমি বিয়ে করব না, কে আমাকে বিয়ে করে দেখি..."	
	ET	“I will not marry, who will marry me let’s see...”	
International	GH	জাপানে পর্যটকবাহী জাহাজ ডুবে নিখোঁজ ২৬	**Matched**
	ET	Tourist ship sinks in Japan and goes missing 26	
	RH	জাপানে পর্যটকবাহী জাহাজ ডুবে নিখোঁজ ২৬	
	ET	Tourist ship sinks in Japan and goes missing 26	
Life-Health	GH	নতুন মৃত্যু ৩৭, শনাক্ত ৩০৪৫	**Sentient**
	ET	New deaths 37, detections 3045	
	RH	আক্রান্ত ছাড়ালো ৫৫ হাজার	
	ET	Number of infected has crossed 55 thousand	
Miscellaneous	GH	চার অক্ষরে সন্তানদের নাম!	**Swapped**
	ET	Four-letters in children’s name!	
	RH	সন্তানের নাম চার অক্ষরে!	
	ET	Children’s names are in four-letters!	
National	GH	ঘোড়া প্রতীক না পেয়ে কেঁদে ফেললেন চেয়ারম্যান প্রার্থী	**Deleted**
	ET	Chairman candidate cried after not getting the horse symbol	
	RH	ঘোড়া প্রতীক না পেয়ে কেঁদে ফেললেন সেই চেয়ারম্যান প্রার্থী	
	ET	That chairman candidate cried after not getting the horse symbol	
Nature	GH	দেশের ৩ বিভাগে বৃষ্টির পূর্বাভাস	**Paraphrased**

	ET	Rain forecast in 3 divisions of the country	
	RH	রাতে বাড়বে তাপমাত্রা, ৩ বিভাগে বৃষ্টির আভাস	
	ET	The temperature will increase at night, there is a chance of rain in 3 divisions	
Opinion	GH	ই-কমার্সবান্ধব বাজেট চাই	**Deleted**
	ET	Want e-commerce friendly budgeting	
	RH	ই-কমার্সবান্ধব বাজেট প্রণয়নে কাজ করতে চাই	
	ET	Want to work on e-commerce friendly budgeting	
Politics	GH	বিদ্যুৎ-গ্যাসের মূল্যবৃদ্ধির সিদ্ধান্ত গণবিরোধী পদক্ষেপ : গণফোরাম	**Inserted**
	ET	The decision to increase the price of electricity and gas is an anti-people move : Public Forum	
	RH	"বিদ্যুৎ ও গ্যাসের মূল্যবৃদ্ধির সিদ্ধান্ত হবে গণবিরোধী"	
	ET	"The decision to increase the price of electricity and gas will be anti-people"	
Religion	GH	পবিত্র শবে বরাত পালিত	**Sentient**
	ET	Holy Shab-e-barat is celebrated	
	RH	ইবাদতে মশগুল ধর্মপ্রাণ মুসলমানেরা	
	ET	Devoted Muslims engaged in prayer	
Science-Tech	GH	মহাকাশে স্যাটেলাইটের সংখ্যা বাড়াচ্ছে ওয়ানওয়েব	**Inserted**
	ET	OneWeb is increasing number of satellites into space	
	RH	মহাকাশে স্যাটেলাইট বাড়াচ্ছে ওয়ানওয়েব	
	ET	OneWeb is increasing satellites into space	
Sports	GH	মেসি এখন পিএসজির জার্সিতে	**Paraphrased**
	ET	Messi is now in PSG jersey	
	RH	পিএসজির হয়ে কবে মাঠে নামছেন মেসি?	
	ET	When is Messi on the field for PSG?	

Table 9: High quality headlines generated on **Shironaam (test)** corpus across all the categories. Here, "Type" means the how the generated headlines are different from the references. We categorize the differences into 5 types: **Inserted** (only one/some word(s) is/are added to reference headline), **Matched** (generated exactly the same), **Swapped** (the only difference is made by swapping one/some word(s) within the reference headline), **Deleted** (the output is about similar to the reference with one/some word(s) less), **Sentient** (generated headline is completely different but a potential competitor against the reference one), **Paraphrased** (paraphrased version of the reference headline). The colored words (*i.e.* teal for *Inserted*, cyan for *Swapped*, brown for *Sentient*, magenta for *Deleted*, and violet for *Paraphrased*) indicate the exact positions where the generated ones are different from the references and no color refers to no change. The generated and reference Bengali headlines, and their corresponding English version are denoted by **GH, RH,** and **ET** respectively.